FlexiTex: Enhancing Texture Generation via Visual Guidance

**Dadong Jiang[1,2], Xianghui Yang[2], Zibo Zhao[2], Sheng Zhang[2], Jiaao Yu[2], Zeqiang Lai[2],
Shaoxiong Yang[2], Chunchao Guo[2*], Xiaobo Zhou[1†], Zhihui Ke[1]**

[1]Tianjin University
[2]Tencent Hunyuan
{patrickdd, xiaobo.zhou}@tju.edu.cn, {seanxhyang, zibozhao, chunchaoguo}@tencent.com

Abstract

Recent texture generation methods achieve impressive results due to the powerful generative prior they leverage from large-scale text-to-image diffusion models. However, abstract textual prompts are limited in providing global textural or shape information, which results in the texture generation methods producing blurry or inconsistent patterns. To tackle this, we present FlexiTex, embedding rich information via visual guidance to generate a high-quality texture. The core of FlexiTex is the Visual Guidance Enhancement module, which incorporates more specific information from visual guidance to reduce ambiguity in the text prompt and preserve high-frequency details. To further enhance the visual guidance, we introduce a Direction-Aware Adaptation module that automatically designs direction prompts based on different camera poses, avoiding the Janus problem and maintaining semantically global consistency. Benefiting from the visual guidance, FlexiTex produces quantitatively and qualitatively sound results, demonstrating its potential to advance texture generation for real-world applications.

Project — https://patrickddj.github.io/FlexiTex/

Introduction

The field of computer graphics aims to enhance the interaction with 3D assets, but the scarcity of high-quality 3D assets is a significant challenge. The creation of such assets requires extensive artistic skills and is a labor-intensive process. However, recent advancements in deep generative models introduce a new paradigm known as Artificial Intelligence Generated Content (AIGC), revolutionizing the creation and manipulation of 3D models (Hong et al. 2023; Chen et al. 2024). Among the various aspects of asset generation, texture generation (Chen et al. 2023a; Metzer et al. 2023) plays a crucial role in adding expressiveness to shapes, and finding applications in industries such as AR/VR, film, and gaming. Furthermore, texture generation holds the potential to overcome the limitations associated with collecting high-quality textured 3D assets.

Recent diffusion models show remarkable advancements in image generation. Noticing their potential on 3D tasks, researchers utilize them to generate textured 3D assets by optimization (Metzer et al. 2023; Poole et al. 2022) and inpainting (Chen et al. 2023a; Zeng et al. 2024), where they render 3D geometry into multiple views and update each view using rich 2D priors, However, these methods neglect cross-view correspondence in 3D space, leaving style inconsistency and seams. Recent works (Zhang et al. 2024a; Liu et al. 2023, 2024; Gao et al. 2024) keep the style consistent and reduce the seams by utilizing batch inference and synchronize multi-view denoising on a shared texture, while still leaving heavy cross-view inconsistency. We argue that it is hard to keep cross-view alignment relying on ambiguous text prompts solely, and frequent feature aggregation introduces a serious variance bias (Liu et al. 2024), resulting in over-smoothness. Moreover, they suffer from the **Janus problem**, due to lack of geometry-aware ability and the bias introduced by ambiguous text descriptions.

Noticing these, we propose a simple yet effective framework to achieve high-quality, precise texture generation, named FlexiTex, which supports text and image conditions both. **First**, we mitigate the ambiguity of text descriptions by converting them to more explicit modalities, *i.e.*, image, which serves as a guide during batch inference and specifies the target object more accurately. Specifically, we introduce a Visual Guidance Enhancement module upon diffusion models to convert the text into an image and inject such a more specific informative guide into the batch inference process through cross-attention. By this, we ensure a more consistent direction during denoising and prevent variance degradation in joint sampling on the shared latent texture. **Second**, we apply a Direction-Aware Adaptation module to alleviate the Janus problem, where we inject direction prompts under different views into the model. This module makes the model more sensitive to direction information and encourages semantic alignment between views. **Third**, the proposed method is not only training-free but also more flexible than previous methods, as it accepts image prompts straightforwardly, by which it can finish texture transfer. This adaptability makes our framework suitable for various geometries and provides more diversity in the generated textures. We conduct comprehensive studies and analyses involving numerous 3D objects from various sources to

---

*Corresponding author.

†Corresponding author.

Figure 1: A collection of textured results generated by FlexiTex. We can produce high-fidelity textures across various objects.

demonstrate the effectiveness of FlexiTexin texture generation, as in Fig. 1.

We summarize our contributions as follows:

- We present a novel and flexible framework for high-quality and accurate texture generation, FlexiTex, which supports text-conditioned generation, image-conditioned generation, and texture transfer tasks.

- We introduce the Visual Guidance Enhancement module to convert the ambiguous text descriptions into more specific images to specify the target between the plausible results.

- We design the Direction-Aware Adaptation module by providing explicit directional information to alleviate Janus problems, which encourages semantic alignment with contents under different views.

Related Work

Text-to-Image Diffusion Models. Recent years have witnessed the advancements of text-to-image diffusion models (Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Zhang, Rao, and Agrawala 2023; Luo et al. 2023; Casas and Comino-Trinidad 2023) for the impressive generative capability to create high-fidelity images. Specifically, Stable Diffusion (Rombach et al. 2022) incorporates a text encoder from CLIP (Radford et al. 2021) and generates realistic images based on input text prompts. Beyond the text conditioning, Zhang *et al.* (Zhang, Rao, and Agrawala 2023) further enhances the model's capabilities. It allows the denoising network to be conditioned on additional input modalities, such as depth or normal maps. Furthermore, Ye *et al.* (Ye et al. 2023) design the effective and lightweight IP-Adaptor to achieve image prompt capability for the pre-trained text-to-image diffusion models. It can be inserted into the current framework and achieve multi-modal image generation. In our work, we leverage ControlNet and Stable Diffusion to offer geometrically conditioned image priors, and IP-Adaptor to support image guidance.

Text-to-texture Synthesis. Generating textures on empty models is a challenging task, which requires precise alignment with geometry and semantic coherence. Early works (Chen, Yin, and Fidler 2022; Siddiqui et al. 2022; Yu et al. 2023; Gao et al. 2022; Chan et al. 2021; Mitchel, Esteves, and Makadia 2024) aim to utilize geometric prior for texture generation. They inject positional information and train geometry-aware generative models from scratch. However, these methods cannot generalize well on various categories and show blurry textures, due to the limited 3D data with high-quality textures for training.

Optimization-based methods (Poole et al. 2022; Metzer et al. 2023; Pan et al. 2024; Youwang, Oh, and Pons-Moll 2024; Chen et al. 2023b; Metzer et al. 2023; Pan et al. 2024; Zhang et al. 2024c) utilize Score Distillation Sampling (SDS) on pre-trained diffusion models to update textures iteratively. Inpainting-based methods (Chen et al. 2023a; Richardson et al. 2023; Cao et al. 2023; Zeng et al. 2024; Zhang et al. 2023; Tang et al. 2024; Chen et al. 2023a; Cao et al. 2023; Ahn et al. 2024) design a sequential inpainting strategy, which fills blank areas of current views based on existing contents from neighbor views. Other methods (Zhang et al. 2024b; Bensadoun et al. 2024; Deng et al. 2024) finetune a multi-view diffusion model to generate a 2×2 gird for multi-view consistency, while they are trained on synthetic datasets and waste rich prior gained from large-scale real image datasets, showing poor diversity.

Recent studies focus on synchronization-based methods (Liu et al. 2023, 2024; Gao et al. 2024; Zhang et al. 2024a; Huo et al. 2024), as they claim that all views contribute equally to texture generation. SyncMVD (Liu et al. 2023) firstly uses a shared latent texture to force consistent latent features from multiple views during denoising. TexPainter (Zhang et al. 2024a) decodes all latent views and performs differentiable inverse rendering at each denois-

ing step, aiming for a texture of higher resolution. GenesisTex (Gao et al. 2024) maintains N unique textures with 1 shared texture to dynamically align the blending weight, and VCD-Texture (Liu et al. 2024) designs 3D-2D co-denoising to strengthen variance.

However, optimization-based methods are time-consuming and suffer from over-saturation, and inpainting-based methods cannot maintain long-range consistency on the whole surface, due to the ambiguity of the text. Among synchronization-based methods, SyncMVD leads to disaturated or over-smoothed results, TexPainter introduces too much noise, and GenesisTex and VCD-Texture require significant memory and computational cost for cross-view attention. In contrast, our FlexiTexleverages visual guidance to generate textures, which not only excels in producing high-quality textures with rich details but also achieves fast and efficient generation.

Image-to-texture Synthesis. Image-to-texture methods (Richardson et al. 2023; Zeng et al. 2024; Pan et al. 2024; Yeh et al. 2024; Chen et al. 2023b) take the image as the user input to generate textures. TEXTure (Richardson et al. 2023) finetune DreamBooth LoRA (Ruiz et al. 2022) and uses the finetuned model as the updated base model for texture inpainting. Optimization-based methods PGC-3D (Pan et al. 2024), Fantastic3D (Chen et al. 2023b) and TextureDreamer (Yeh et al. 2024) apply SDS conditioned on input image prompts for appearance modeling. These methods require a long time for finetuning or repainting, and cannot transfer the semantic identity well to target meshes. Paint3D (Zeng et al. 2024) applies image prompts directly on UV generation in the refinement stage, but cannot generalize well on complex UV maps, usually leaving certain artifacts on final textures. Unlike these previous methods, our method can support high-quality texture generation in a training-free manner, which is flexible because it requires no additional adjustment for image prompts.

Method

FlexiTexis designed to generate high-quality texture maps given an untextured mesh conditioned on either text or image. The overview is shown in Fig. 2. In this section, we first provide the backgrounds of diffusion models, mesh rendering, and texture warping. Following this, we introduce the Visual Guidance Enhancement module, which is designed to overcome issues related to over-smoothing. Then, we present the Direction-Aware Adaptation module, which is specifically designed to incorporate direction information during texture generation, thereby addressing the Janus problem.

Preliminary

We introduce the preliminaries here, including diffusion models, mesh rendering, and texture warping.

Stable Diffusion & Controlnet. The diffusion model (Ho, Jain, and Abbeel 2020) consists of a forward process $q(\cdot)$ and a reverse denoising process $p_\theta(\cdot)$. The forward process progressively corrupts the original data, denoted as x_0, with noise $\epsilon \sim \mathcal{N}(0, 1)$ to a noisy sequence $x_1, \ldots, x_T$, follow-

ing a Markov chain as:

$$q(x_t|x_{t-1}, y) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I),$$
$$x_t = \sqrt{\beta_t}x_0 + \sqrt{1 - \beta_t} \cdot \epsilon, \quad (1)$$

where β_t represents the variance schedule for $t = 1, \ldots, T$; y is the condition. The reverse process denoises the pure data from x_T as:

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, y), \Sigma_\theta(x_t, t, y)), \quad (2)$$

where μ_θ and Σ_θ denote the mean and variance predictions, respectively, which are obtained from a trainable network parameterized by θ.

Besides, we introduce ControlNet (Zhang, Rao, and Agrawala 2023) which injects low-level control, such as depth map d, during the denoising process of Stable Diffusion. The predicted noise of the U-Net with ControlNet is represented as $\epsilon_\theta(x_t, y, d, t)$

Texture warping. Given a mesh $\mathcal{M}$, a texture map $\mathcal{T}$ and a viewpoint C, we can use the rasterization function $\mathcal{R}$ to render an image. After rasterization, each valid pixel on rendered images corresponds to one on texture. However, these pixels are scattered in UV space after back projecting, necessitating Voronoi filling (Aurenhammer 1991) to fill up all blank regions in texture warping ($\mathcal{W}^{-1}$). Given $\{\mathbf{s}_i | \mathbf{s}_i = (u_i, v_i)\}$ represents the UV coordinates of the i-th latent texel, we first generate the Voronoi diagram partitions the domain $\mathcal{T}$ into a set of regions $\{\mathcal{V}_i\}$, where each region $\mathcal{V}_i$ is defined as:

$$\mathcal{V}_i = \{\mathbf{p} \in \mathcal{T} \mid \|\mathbf{p} - \mathbf{s}_i\| \leq \|\mathbf{p} - \mathbf{s}_j\|, \forall j \neq i\}. \quad (3)$$

Here, $\mathbf{p}$ is the 2D point position and $\| \cdot \|$ denotes the Euclidean distance. Then we define a procedural function $\mathcal{P}_i : \mathcal{V}_i \to \mathbb{R}^3$ that generates the filling for the region $\mathcal{V}_i$ based on the texels $\mathbf{s}_i$ and the position $\mathbf{p}$ within the region. Finally, we record visible masks on the texture map and apply specific boundary conditions near the edges to ensure a smooth and consistent filling. The final texture is represented as the union of the filled regions: $\mathcal{T} = \bigcup_{i=1}^{N} V_i$, where N is the number of regions.

Visual Guidance Enhancement

In FlexiTex, we design a Visual Guidance Enhancement module to align the denoising inference on multiple views. Explicitly, for text input $\boldsymbol{x}^{text}$, we first utilize a text-to-image model to generate image prompt $\boldsymbol{x}_{img}$. Its corresponding semantic information $\boldsymbol{c}_{img}$ is injected into the denoising process in a cross-attention manner through IP-Adaptor (Ye et al. 2023). To be specific, given the image features $\boldsymbol{c}_{img}$, the output of new cross-attention $\mathbf{Z}'$ is computed as follows:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}')$$
$$= \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}', \quad (4)$$

where, $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K}' = \boldsymbol{c}_{img}\mathbf{W}'_k$ and $\mathbf{V}' = \boldsymbol{c}_{img}\mathbf{W}'_v$ are the query, key, and values matrices from the image features.
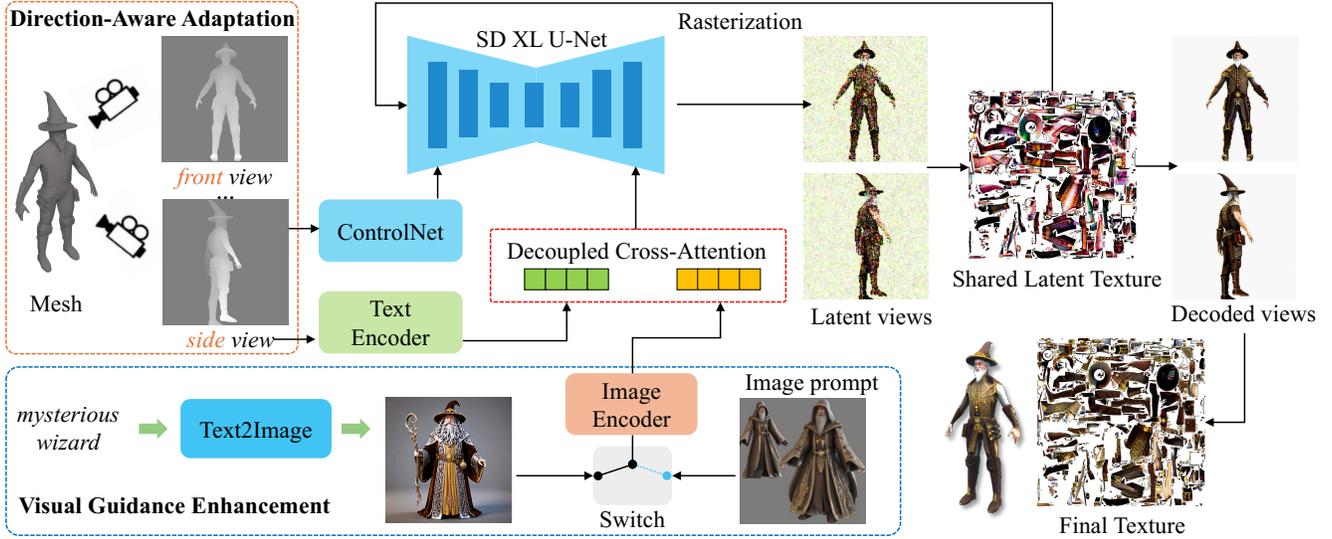
Figure 2: The framework of FlexiTex. Given a mesh with corrected orientation and text/image prompts, the Visual Guidance Enhancement module extracts **visual features** ■ to provide informative guidance during the denoising steps. The Direction-Aware Adaptation module then extracts **direction features** ■ according to camera poses. Visual and direction features are integrated into the pre-trained UNet model with decoupled cross-attention, along with depth maps. Through iterative texture warping and rasterization, our method generates high-fidelity textures of both rich details and multi-view consistency.

Since images provide more specific information, the image-guided batch diffusion process also benefits from this, in which case the denoising direction is more consistent. With intermediate latent features which are normalized by explicit visual features, there is no longer a huge variance degradation in joint sampling on shared latent texture, compared with the text-guided way. Images also provide clear guidance on appearance modeling, prompting a similar style on multiple views and maintaining global consistency.

In addition, our framework supports image-to-texture tasks. We do not require the image prompt to be totally aligned with mesh geometry but consider it a style-transferring task to maintain the most significant input semantic information on the target mesh.

Direction-Aware Adaptation

We introduce Direction-Aware Adaptation which utilizes a Text Encoder with U-Net to provide spatial information and solve the Janus problem. We find that SD-XL is more sensitive to direction prompts. When we add prompts *i.e.*, *front/side/back view*, it can understand these well and generate more semantically aligned results. Noticing that the orientation of the original shapes is not guaranteed, we first correct their orientation to ensure they are front-facing. This pre-processing step defines the standard orientation and determines the writing of subsequent direction prompts. Then we generate prompt $\{c_{view,i}\}_{i=1}^{N}$ according to the elevation and azimuth for each view, organized as "from $<?>$ view". Given such direction prompts, the denoising process for $view_i$ is conditioned on visual features c_{img} and direction features c_{view} via cross-attention, written as:

$$\mathbf{Z}'' = \mathrm{Softmax}(\frac{\mathbf{QK}^{\top}}{\sqrt{d}})\mathbf{V} + \mathrm{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^{\top}}{\sqrt{d}})\mathbf{V}', \quad (5)$$

where $\mathbf{Q} = \mathbf{ZW}_q, \mathbf{K} = c_{view}\mathbf{W}_k, \mathbf{V} = c_{view}\mathbf{W}_v, \mathbf{K}' = c_{img}\mathbf{W}'_k, \mathbf{V}' = c_{img}\mathbf{W}'_v$. In our framework, the image prompt provides visual guidance on appearance and text description provides direction information, by which the direction-aware single-view generative model can naturally become a multi-view generator during batch inference and synchronized sampling, as shown in Fig. 2.

Experiments

Setup

Implementation Details. Our experiments are conducted on an NVIDIA A100 GPU. For denoising epoch, we use DDIM (Song, Meng, and Ermon 2020) as the sampler. We set the number of iterations to 30 steps, the CFG scale (classifier-free guidance scale) for Direction-Aware Adaptation to 12, and the scale of Visual Guidance Enhancement to 0.6. Texture warping for latent views is used in the first 24 steps. We sample 8 views for a mesh, and the elevations and azimuths are $(-180°, 15°), (-120°, -15°), (-60°, 15°), (0°, -15°), (60°, 15°), (120°, -15°), (-180°, -45°), (0°, 45°)$. We implement the rendering function by Pytorch3D (Ravi et al. 2020; Paszke et al. 2017).

Dataset. We collect 60 meshes with corresponding text prompts and 60 meshes with corresponding image prompts to evaluate the text-to-texture and image-to-texture generation ability, respectively. These meshed are randomly sampled from Objaverse (Deitke et al. 2022), Objaverse-XL (Deitke et al. 2023) and ShapeNet (Chang et al. 2015).
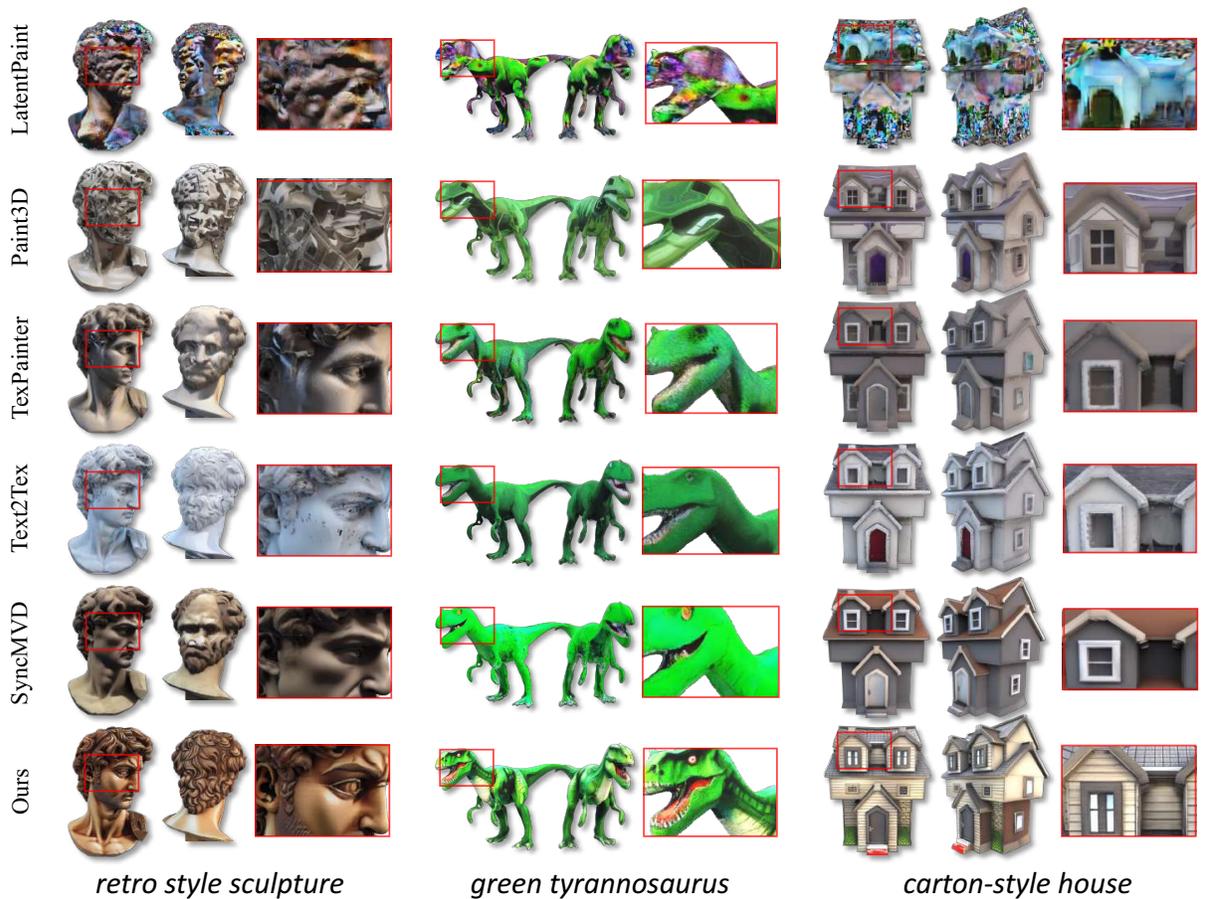
Figure 3: Qualitative comparisons on texture generation conditioned on text prompt. While previous methods result in an over-smoothed appearance and global inconsistencies such as the Janus problem, our method generates globally consistent and higher-quality textures rich in detail (refer to supplements for more results).

Evaluation metrics. Fréchet inception distance (FID) and Kernel Inception Distance (KID) measure the feature dissimilarity between two image collections, with feature extraction performed using the Inception V3 (Szegedy et al. 2016), as metrics for the realism and diversity. Moreover, we also utilize the CLIP Score metric (Zhengwentai 2023) to assess the alignment with original inputs.

Baselines. For text-to-texture generation, we compare FlexiTexwith inpainting-based methods **Text2Tex** (Chen et al. 2023a), **Paint3D** (Zeng et al. 2024), optimization-based methods **Latent-Paint** (Metzer et al. 2023), synchronization-based methods **TexPainter** (Zhang et al. 2024a), **SyncMVD** (Liu et al. 2023). For image-to-texture generation, we compare our FlexiTexwith inpainting-based methods **TEXTure** (Richardson et al. 2023), **Paint3D** (Zeng et al. 2024), optimization-based methods **PGC-3D** (Pan et al. 2024). For a fair comparison, we re-run their official codes using the same meshes and prompts.

Quantitative Analysis

FID & KID. Following GenesisTex (Gao et al. 2024), we use samples from pre-trained diffusion models as ground truth labels. We render depth maps from 16 cameras around the mesh as ControlNet inputs and we maintain the same prompts for inference. We modify the ground truth images' background to white to ensure the focus remains on the texture. We render each mesh from the same views, assessing quality via FID and KID. Tab. 1 shows our method outperforms baselines, indicating superior synthesis quality and closer alignment to ground truth, attributed to the explicit information provided by visual guidance.

ClipScore. For text-to-texture tasks, the CLIP score is derived by comparing rendered views with text prompts, while for image-to-texture tasks, it is computed by comparing rendered views with image prompts. Semantic consistency deviations may occur in text-to-texture tasks due to the text-to-image module, resulting in a slightly lower CLIP Score than TexPainter, as shown in Tab. 1. However, our method enhances texture quality and realism. For image-to-texture tasks, Paint3D improves alignment by injecting image features into UV map refinement, but struggles with complex UV maps due to semantic differences. Conversely, Flexi-Tex achieves the highest CLIP Score by employing visual guidance on multi-view inference, ensuring semantic con-

Figure 4: Qualitative comparisons on texture generation conditioned on image prompt. Compared with baselines, FlexiTex achieves better alignment with input images on target meshes. Our results are also superior in texture integrity and quality.

(a) Quantitative comparisons conditioned on text prompt.

	FID ↓	KID ↓	CLIP Score ↑	Time ↓
Latent-Paint	137.56	44.09	31.32	8 min
Paint3D	99.05	13.28	33.44	3.1 min
TexPainter	96.97	13.51	**34.48**	40 min
Text2Tex	90.98	10.62	34.33	6 min
SyncMVD	87.65	9.53	34.30	**1.8 min**
Ours	**81.72**	**8.78**	34.31	2.2 min

(b) Quantitative comparisons conditioned on image prompt.

	FID ↓	KID ↓	CLIP Score ↑	Time ↓
TEXTure	141.26	51.11	82.92	32 min
PGC-3D	133.76	32.37	79.05	19 min
Paint3D	103.44	18.71	87.43	3.2 min
Ours	**82.52**	**11.97**	**89.99**	**2.4 min**

Table 1: Quantitative comparisons with baselines. ↓ and ↑ are used to indicate the performance in relation to the score. ↓ indicates better performance with a lower score, while ↑ indicates better performance with a higher score.

sistency with input images.

Speed. Compared with optimization-based and inpainting-based methods requiring sequential sampling, our approach simultaneously generates multiple views once. Rapid texture warping further accelerates generation, in contrast to TexPainter using 40-minute due to enforced differentiable rendering in each denoising step.

Qualitative Analysis

As presented in Fig. 3 and Fig. 4, we qualitatively evaluate the quality of the generated texture conditioned on text and image, respectively. The Latent-Paint method produces textures that are broken by noise and impurities, as a consequence of the inherent limitation of SDS. Both Paint3D and Text2Tex exhibit noticeable blurriness at texture seams and suffer from the multi-face issue, thereby failing to maintain multi-view consistency. TexPainter and SyncMVD struggle to preserve clarity on texture, ending up with an over-smoothed and monotonous appearance. Through Visual Guidance Enhancement, FlexiTex overcomes the over-smooth problem and preserves more high-frequency details during generation. Direction-Aware adaptation also strengthens geometric alignment on side or back views, alleviating the Janus problem.

For image-to-texture tasks shown in Fig. 4, both TEXTure and PGC-3D struggle to retain the semantic information from image prompts, resulting in low-quality, noisy textures. Paint3D, while partially preserving semantic information, generates a significant number of texture fragments. In contrast, we implement a decoupled cross-attention strategy from visual features and direction features, achieving semantic alignment with the style of image prompts on target meshes and exhibiting a vibrant appearance.

Ablation Studies

Effectiveness of Visual Guidance Enhancement. To examine the influence of various prompts on texture quality, we compare three types: **simple prompts** comprising no more than five words, **refined prompts** expanded by a Large Language Model (LLM) Llama-3 (AI@Meta 2024), and **Visual Guidance Enhancement** using our method. As depicted in Fig. 5, textures derived from simple prompts can appear blurry or desaturated. While refined prompts marginally slightly address this issue, their results still lack vibrant details. However, with Visual Guidance Enhancement, we can observe a significant improvement in detail
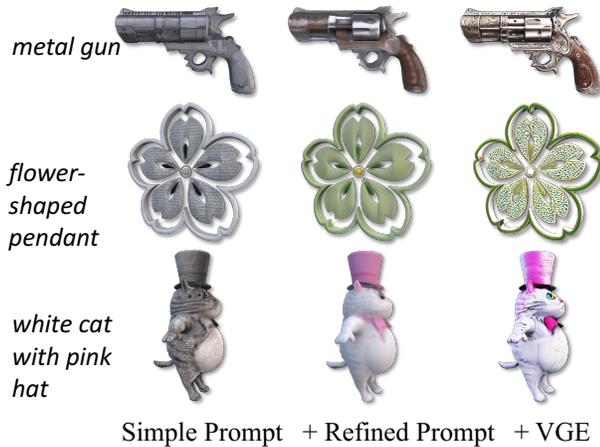
Figure 5: Ablation results on Visual Guidance Enhancement.

Figure 6: Ablation results on Visual Guidance Enhancement.

richness. Similar to VCD-Texture (Liu et al. 2024), we visualize the standard deviation (std) curve of the latent features from foreground parts during the denoising phase, in Table 7. Here, Visual Guidance Enhancement achieves the highest feature variance, indicating high-frequency details. Apart from the rich details brought by VGA, we also achieve better **style consistency** in Fig. 6. For example, for prompt *'a carton-style house'*, the original results can show grey on one side and orange on the other side, while with the VGA module, these two sides show consistent color.

Effectiveness of Direction-Aware Adaptation. We visualize the effects of Direction-Aware Adaptation (DAA) on alleviating multi-face problem in Fig. 8. With reinforcement of specific direction on target meshes, the contents of each view are correct(e.g., a lizard shows one eye from the side view, the front of the iPhone is a screen, and a human only has one face). We also experiment on 200 human cases, 200 animal cases, and 200 object cases. Tab. 2 shows a lower multi-face percentage.

Conclusion

In this paper, we introduce FlexiTex, a novel framework designed for high-fidelity texture generation on 3D objects, accommodating both text and image prompts. We also mitigate multi-face issues across various types of objects. Experiments demonstrate that FlexiTex is superior to baseline methods in both quantitative and qualitative measures. However, our method does have certain limitations. Specifically, our generated results have not yet decoupled lighting information, leading to potential artifacts in highlights or shadows. Additionally, in areas where multiple views are inconsistent or unobserved, minor dirty spots may appear. Future research could explore material generation for re-lighting tasks and improved texture warping strategies, such as maintaining a color field for smooth transitions in conflict areas.
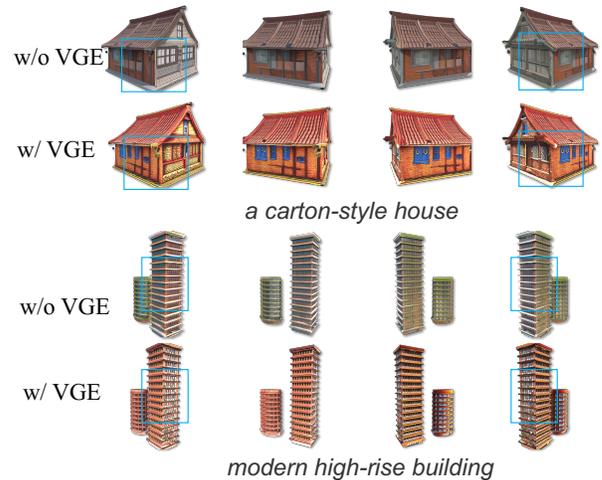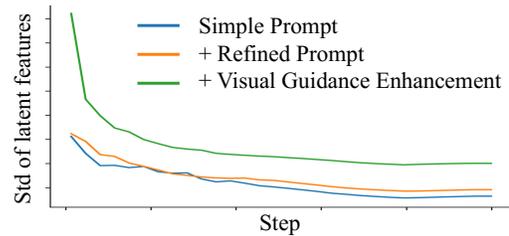

Figure 7: Ablation results of standard deviation (std) on Visual Guidance Enhancement.

Figure 8: Ablation results on Direction-Aware Adaptation. We show results from **side** and **front** views.

	Animal ↓	Object ↓	Human ↓
w/o DAA	48.00 %	8.00 %	24.00 %
w/ DAA	**21.50 %**	**3.00 %**	**16.50 %**

Table 2: Ablation results on Direction-Aware Adaptation (DAA). We calculate multi-face percentages on 200 human cases, 200 animal cases, and 200 object cases.

Acknowledgments

References

Ahn, J.; Cho, S.; Jung, H.; Hong, K.; Ban, S.; and Jung, M.-R. 2024. ConTEXTure: Consistent Multiview Images to Texture. *arXiv preprint arXiv:2407.10558*.

AI@Meta. 2024. Llama 3 Model Card.

Aurenhammer, F. 1991. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3): 345–405.

Bensadoun, R.; Kleiman, Y.; Azuri, I.; Harosh, O.; Vedaldi, A.; Neverova, N.; and Gafni, O. 2024. Meta 3D Texture-Gen: Fast and Consistent Texture Generation for 3D Objects. *arXiv preprint arXiv:2407.02430*.

Cao, T.; Kreis, K.; Fidler, S.; Sharp, N.; and Yin, K. 2023. TexFusion: Synthesizing 3D Textures with Text-Guided Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Casas, D.; and Comino-Trinidad, M. 2023. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; Mello, S. D.; Gallo, O.; Guibas, L.; Tremblay, J.; Khamis, S.; Karras, T.; and Wetzstein, G. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, D. Z.; Siddiqui, Y.; Lee, H.-Y.; Tulyakov, S.; and Nießner, M. 2023a. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18558–18568.

Chen, H.; Shi, R.; Liu, Y.; Shen, B.; Gu, J.; Wetzstein, G.; Su, H.; and Guibas, L. 2024. Generic 3D Diffusion Adapter Using Controlled Multi-View Editing. arXiv:2403.12032.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023b. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22246–22256.

Chen, Z.; Yin, K.; and Fidler, S. 2022. AUV-Net: Learning Aligned UV Maps for Texture Transfer and Synthesis. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv preprint arXiv:2212.08051*.

Deng, K.; Omernick, T.; Weiss, A.; Ramanan, D.; Zhu, J.-Y.; Zhou, T.; and Agrawala, M. 2024. FlashTex: Fast Relightable Mesh Texturing with LightControlNet. In *European Conference on Computer Vision (ECCV)*.

Gao, C.; Jiang, B.; Li, X.; Zhang, Y.; and Yu, Q. 2024. GenesisTex: Adapting Image Denoising Diffusion to Texture Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4620–4629.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. In *Advances In Neural Information Processing Systems*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Huo, D.; Guo, Z.; Zuo, X.; Shi, Z.; Lu, J.; Dai, P.; Xu, S.; Cheng, L.; and Yang, Y.-H. 2024. TexGen: Text-Guided 3D Texture Generation with Multi-view Sampling and Resampling. *ECCV*.

Liu, S.; Yu, C.; Cao, C.; Qian, W.; and Wang, F. 2024. VCD-Texture: Variance Alignment based 3D-2D Co-Denoising for Text-Guided Texturing. *arXiv preprint arXiv:2407.04461*.

Liu, Y.; Xie, M.; Liu, H.; and Wong, T.-T. 2023. Text-Guided Texturing by Synchronized Multi-View Diffusion. *arXiv preprint arXiv:2311.12891*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv:2310.04378.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12663–12673.

Mitchel, T. W.; Esteves, C.; and Makadia, A. 2024. Single Mesh Diffusion Models with Field Latents for Texture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7953–7963.

Pan, Z.; Lu, J.; Zhu, X.; and Zhang, L. 2024. Enhancing High-Resolution 3D Generation through Pixel-wise Gradient Clipping. In *International Conference on Learning Representations (ICLR)*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501*.

Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; and Cohen-Or, D. 2023. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, 1–11.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Siddiqui, Y.; Thies, J.; Ma, F.; Shan, Q.; Nießner, M.; and Dai, A. 2022. Texturify: Generating Textures on 3D Shape Surfaces. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, volume 13663 of *Lecture Notes in Computer Science*, 72–88. Springer.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tang, J.; Lu, R.; Chen, X.; Wen, X.; Zeng, G.; and Liu, Z. 2024. InTeX: Interactive Text-to-texture Synthesis via Unified Depth-aware Inpainting. *arXiv preprint arXiv:2403.11878*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yeh, Y.-Y.; Huang, J.-B.; Kim, C.; Xiao, L.; Nguyen-Phuoc, T.; Khan, N.; Zhang, C.; Chandraker, M.; Marshall, C. S.; Dong, Z.; et al. 2024. TextureDreamer: Image-guided Texture Synthesis through Geometry-aware Diffusion. *arXiv preprint arXiv:2401.09416*.

Youwang, K.; Oh, T.-H.; and Pons-Moll, G. 2024. Paint-it: Text-to-Texture Synthesis via Deep Convolutional Texture Map Optimization and Physically-Based Rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, X.; Dai, P.; Li, W.; Ma, L.; Liu, Z.; and Qi, X. 2023. Texture Generation on 3D Meshes with Point-UV Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4206–4216.

Zeng, X.; Chen, X.; Qi, Z.; Liu, W.; Zhao, Z.; Wang, Z.; Fu, B.; Liu, Y.; and Yu, G. 2024. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4252–4262.

Zhang, H.; Pan, Z.; Zhang, C.; Zhu, L.; and Gao, X. 2024a. TexPainter: Generative Mesh Texturing with Multi-view Consistency. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Zhang, J.; Tang, Z.; Pang, Y.; Cheng, X.; Jin, P.; Wei, Y.; Yu, W.; Ning, M.; and Yuan, L. 2023. Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting. *arXiv preprint arXiv:2312.13271*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, L.; Wang, Z.; Zhang, Q.; Qiu, Q.; Pang, A.; Jiang, H.; Yang, W.; Xu, L.; and Yu, J. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 43(4): 1–20.

Zhang, Y.; Liu, Y.; Xie, Z.; Yang, L.; Liu, Z.; Yang, M.; Zhang, R.; Kou, Q.; ; Lin, C.; Wang, W.; and Jin, X. 2024c. DreamMat: High-quality PBR Material Generation with Geometry- and Light-aware Diffusion Models. In *SIGGRAPH*.

Zhengwentai, S. 2023. clip-score: CLIP Score for PyTorch. https://github.com/taited/clip-score. Version 0.1.1.